Lexical Normalization of Spanish Tweets

Jhon Adrián Cerón-Guzmán
jacerong@unal.edu.co

Elizabeth León-Guzmán
eleonguz@unal.edu.co

MIDAS Research Group, Departamento de Ingeniería de Sistemas e Industrial
Universidad Nacional de Colombia, Bogotá D.C., Colombia

ABSTRACT

Twitter data have brought new opportunities to know what happens in the world in real-time, and conduct studies on the human subjectivity on a diversity of issues and topics at large scale, which would not be feasible using traditional methods. However, as well as these data represent a valuable source, a vast amount of noise can be found in them. Because of the brevity of texts and the widespread use of mobile devices, non-standard word forms abound in tweets, which degrade the performance of Natural Language Processing tools. In this paper, a lexical normalization system of tweets written in Spanish is presented. The system suggests normalization candidates for out-of-vocabulary (OOV) words based on similarity of graphemes or phonemes. Using contextual information, the best correction candidate for a word is selected. Experimental results show that the system correctly detects OOV words and the most of cases suggests the proper corrections. Together with this, results indicate a room for improvement in the correction candidate selection. Compared with other methods, the overall performance of the system is above-average and competitive to different approaches in the literature.

Keywords

Finite-state transducers, language modeling, lexical normalization, out-of-vocabulary words, Spanish tweets, Twitter

1. INTRODUCTION

Social media has changed the paradigm of information generation and consumption. In platforms such as Twitter[1], users generate content on real-world events at the same time they occur, e.g., real-time reports on natural disasters, or share their views on a diversity of issues and topics often intended to impact other users' or companies' decisions, e.g., opinions about a product that motivate its purchase or improvement [12]. Thus, Twitter data constitute an useful source to gain subjective/objective insight on different matters. The literature presents applications to get benefit from such data including event detection and analysis [11], sentiment analysis [4], and even predictions [22]. However, because of the brevity of the tweets[2] and the widespread use of mobile devices [24], Twitter is also a rich source of noisy data [7] containing many non-standard word forms. That is why several lexical variation phenomena that occur on the content generation, need to be tackled within the pipeline of a Natural Language Processing (NLP) task, in order to improve the quality of natural language analysis [5, 7].

The language used in Twitter is mainly characterized by an informal genre and a free writing style [5], where initialisms, shortenings, homophonic confusion, character repetition, and misuse of uppercase are commonly used to either save characters or denote emphasis in tweets. To deal with these lexical variation phenomena, two lines of research have been proposed in the literature: the first one has to do with the development of NLP tools that adapt to short and noisy texts [13]; the second one proposes as a preprocessing step within the pipeline of a NLP task, the normalization of non-standard word forms to their standard lexical forms [2, 5, 7]. This second approach has received more attention from the scholars, showing promising results.

Initial lexical normalization approaches of tweets have focused on English [7, 8]. However, Twitter content in languages such as Spanish rapidly increases [23], for which normalization strategies to deal with lexical variation phenomena present in Spanish tweets becomes a major issue in order to boost NLP applications that exploit user-generated content in that language.

In this paper, a lexical normalization system of Spanish tweets is presented. The overall process of lexical normalization follows a sequential approach that goes from the detection of out-of-vocabulary (OOV) words in a tweet, to the correction candidate selection for a word from a set of normalization proposals. In contrast to [7], where a one-to-one normalization approach was developed, this is, one OOV word is normalized to one standard lexical form, this work proposes a one-to-many normalization approach to also deal with word segmentation problems such as lack of spacing between words.

The remainder of this paper is organized as follows: Section 2 revises related work on lexical normalization of tweets. The system architecture, which is divided into three compo-

[1] `https://twitter.com/`

[2] User posts on Twitter are known as tweets, which have a 140-character limit.

WWW'16 Companion, April 11–15, 2016, Montréal, Québec, Canada.
ACM 978-1-4503-4144-8/16/04.
http://dx.doi.org/10.1145/2872518.2890558.

nents and considers a post-processing step, as well as the set of lexical resources employed by the system to suggest normalization candidates for OOV words, may be read in Sections 3 and 4, respectively. In Section 5 the experimental development of the system and the evaluation of its performance are presented. Finally, conclusions are drawn in Section 6.

2. RELATED WORK

A large body of literature exists on lexical normalization of tweets written in English. Han and Baldwin [7] characterized the types of OOV words present in tweets, finding that these correspond to a heterogeneous collection of ill-formed words and proper nouns. As a critical step within the process of lexical normalization, they proposed an automatic method to distinguish between correct OOV words and ill-formed OOV words, and for the latter normalization candidates were suggested, from which the best were selected based on contextual inference, dependency features, and string similarity measures.

Han, Cook, and Baldwin [8] proposed a dictionary-based approach to normalize OOV words that fails to adapt to new domains, thus recording low values of recall, and does not normalize complex cases of OOV words with two or more possible standard lexical forms, for which contextual information may be used in order to resolve ambiguities.

A common shortcoming of these cited works is that they both have focused on one-to-one normalization, i.e., one OOV word is normalized to one standard lexical form. However, OOV words may also be normalized by splitting fused words, which is why a one-to-many normalization approach is required.

As an initiative to foster research in the field, the Tweet-Norm 2013 shared task [2] was organized to create a benchmark for lexical normalization of tweets written in Spanish. The resources provided by the organizing committee were used to conduct experiments and evaluate the performance of the system presented in this paper. The highest ranked participating systems are described below.

Porta and Sancho [15] used several weighted finite-state transducers that were applied in cascade to generate the confusion set of each OOV word. The standard lexical forms were suggested by their similarity to the graphemes or phonemes that make an OOV word, and the candidate selection was made by the application of a trigram language model.

Gamallo, Garcia, and Pichel [6] distinguished normalization candidates between primary and secondary variants. The former correspond to candidates that only differ from an OOV word with regard to one of several linguistic phenomena (i.e., uppercase/lowercase confusion, character repetition, or frequent spelling errors); otherwise, secondary variants were generated using the Levenshtein distance. They also used a language model in the candidate selection. Without using contextual information, Ageno et al. [1] selected the normalization candidate from a confusion set generated by a set of expert modules, through a weighted voting scheme. Saralegi and San Vicente [21] assumed that all the named entities recognized by a third-party language analyzer were correct OOV words; however, as it will be proved in this paper, these must be carefully treated because users misuse uppercase in tweets, e.g., to denote emphasis, thereby producing false positive of named entities.

Finally, Cotelo et al. [5] conducted a complete study of the types of OOV words present in Spanish tweets. They proposed a modular architecture for lexical normalization, in which each module addressed a specific error phenomenon. Thus, each module suggested normalization candidates, and the best one was selected through a weighted voting scheme.

3. THE SYSTEM ARCHITECTURE

The overall process of lexical normalization follows a sequential approach that goes from the detection of OOV words in a given tweet, to the correction candidate selection for a word. The approach is structurally divided into three components that are discussed below: in the first one, a third-party language analyzer is used for performing tokenization of tweets and lexical analysis of in-vocabulary (IV) words, while non-standard word forms are detected (i.e., OOV words); the second one generates normalization candidates for each OOV word (i.e., the confusion set); finally, the third one selects the best candidate from the confusion set of each OOV word, taking into account contextual information. After the selection, a post-processing is applied to uppercase the correction for a word when one of several conditions is satisfied.

3.1 Detecting OOV Words

The morphological analyzer of FreeLing [14] is used for detecting OOV words. Once a given tweet has been tokenized, each resulting token is passed through a set of basic modules (e.g., dictionary lookup, suffixes check, detection of numbers and dates, named entity recognition, etc.) for identifying standard word forms and other valid constructions. If a token is not recognized by any of the modules, it is marked as OOV. In this step, specific Twitter terms like user mentions (e.g., @twitter), hashtags (e.g., #twitter), and "RT" (retweet) and other expressions such as URLs are treated as valid constructions.

While some experiments were conducted on a development set, an unexpected behavior of the Named Entity Recognition module of FreeLing was observed.[3] Specifically, tokens starting with a capital letter or completely written in uppercase were mostly wrong recognized as named entities, because the capitalization rules [17] are not taken into account by users that write tweets misusing uppercase, e.g., to denote emphasis, thereby producing false positives of named entities that must be carefully treated. For example, given the tweet *"Lo mejor es que me da igual todo SOI FELIZ"* (`The best is that i do not care anything, I AM HAPPY`), the tokens *"SOI FELIZ"* (`i am happy`) are wrongly recognized as an entity, being *"SOI"* a typo of the standard word form *"soy"* (`i am`) and *"FELIZ"* (`happy`) a standard word form. Therefore, each token recognized as an entity is looked up in the dictionary of standard words, and if there is not an entry matching the token, it is marked as OOV.

3.2 Confusion Set Generation

Once the OOV words have been detected, a first issue to be tackled is to determine if a given OOV is either a correct word that is not in the standard dictionary, or a token requiring to be normalized to its canonical form. That is, it is explicitly necessary to distinguish between correct OOV words and ill-formed OOV words [7]. For the for-

[3]Using the "basic" recognizer.

mer, the OOV itself would remain unchanged, while for the latter, several lexical variation phenomena should be dealt with, including: character repetition (e.g., *claseeeesss* → *clases* (`classes`)) and alteration of valid onomatopoeia (e.g., *ajajajjaja* → *ja*); language-dependent orthographic errors [5,6]: missing of diacritical marks (e.g., *tendre* → *tendré* (`i will have`)), uppercase/lowercase confusion (e.g., *francia* → *Francia* (`France`)), and letter confusion (v → b, ll → y, h → 0); initialisms (e.g., *xk* → *porque* (`because`)), shortenings (e.g., *pa* → *para* (`for`)), and letter omissions [15]; homophonic confusion (e.g., *pokitin* → *poquitín* (`little bit`)) [5] and standard non-correct endings (e.g., *mercao* → *mercado* (`market`)) [15]; and word segmentation problems (e.g., *alomejor* → *"a lo mejor"* (`at best`)) [15]. Thus, in order to determine if an OOV token is a correct word, it is first included in its confusion set; if the OOV token is which best fits a language model, it is then considered as correct. The confusion set generation is discussed below.

A confusion set is generated by either one of two sequential phases. The first one involves a set of simple rules intended to tackle some of the most common lexical variation phenomena present in Spanish tweets. If an OOV word is recognized by one of these rules, its canonical form is provided; otherwise, the second one generates normalization candidates that are identical or similar to the graphemes or phonemes that make the OOV word. During the entire process, the consecutive repetition of a same letter is reduced to one and two occurrences, thus generating three different versions of the OOV word (the first one being the OOV itself, the second one with no letter repetition,[4] and the third one with at most two consecutive repetitions); the repetition reduction is inspired by the approach proposed in [1]. Likewise, the treatment of unknown characters, taking as reference the Spanish alphabet [16], is conducted by representing them to their closest ASCII variant, using the *unidecode*[5] module for the mapping.

The confusion set generation comprises a set of finite-state networks developed to deal with the foregoing lexical variation phenomena. These networks are computationally efficient for tasks such as natural-language morphological analysis, and for their mathematical properties, which are well understood, it is allowed to manipulate and combine them in ways that would be impossible using traditional algorithmic programs [3].

3.2.1 Matching Simple Rules

As discussed above, a set of simple rules was designed to tackle some of the most common lexical variation phenomena present in Spanish tweets, namely: alteration of valid onomatopoeia, missing of diacritical marks, initialisms, and shortenings. These rules are described as regular expressions compiled into finite-state transducers using the Foma library [10]. Thus, if an OOV word is accepted by the language of a network, i.e., a transducer, its canonical form is provided; otherwise, if the OOV word is rejected by the set of transducers, the process of the confusion set generation is applied.

Note that in this phase, two or more target words can be suggested, instead to be directly provided the normalization of an OOV word. For example, let the OOV word be *"siii"*, and the network be the composition of the transducers to deal with the missing of diacritical marks,[6] and accept all the valid Spanish words, the following normalization candidates are returned by the process of generation: *"si"* (`if`) and *"sí"* (`yes`).[7] However, in the most of cases, the normalization of an OOV word is directly provided. In this regard, initialisms and shortenings are dealt with a network whose language consists of frequent OOV words that may be included within these phenomena, and their normalization can be provided unambiguously. The language is a normalization dictionary that was built from the most frequent OOV words observed in a development set, and Internet slang used in Spanish tweets.

3.2.2 Generating the Confusion Set

To tackle the remaining lexical variation phenomena, in this phase a set of normalization candidates is generated. The candidates are elements of the union of the standard dictionary and the gazetteer of proper nouns, which are identical or similar to the graphemes or phonemes that make an OOV word.

Firstly, the OOV word is converted into its phonetic transcription using the International Phonetic Alphabet (IPA). The phonetic transcription makes the IPA phonemes /ʝ/ and /ʎ/ equivalent, which is a phenomenon that occurs in many dialects of the Spanish language [20]. The linguistic phenomenon of *seseo* [18], homophonic confusion, and standard non-correct endings are also modeled by the transducer that makes the transcription; the phenomena of uppercase/lowercase confusion and letter confusion are implicitly modeled. Thus, normalization candidates are suggested by their phonetic similarity to the OOV word. Likewise, a suffixes search is performed to recognize inflected forms that are not found in the standard dictionary, namely: clitics attached to verbal forms of infinitive, imperative, and gerund, i.e., enclitic pronouns; adverbs ending in *-mente*; and diminutive forms of adjectives, adverbs and nouns [25]. Therefore, if the OOV is recognized as an inflected word form, it is suggested as a candidate with the proper accentuation.

Secondly, if no candidates are generated by the above approach, in this one the most complex cases of OOV words, mainly characterized by the phenomena of letter omissions and word segmentation problems, are tackled. To deal with the first phenomenon, a transducer inserts only one vowel in any position of the OOV word, as it was proposed in [15]. Inspired by [15] and [1], the second phenomenon is dealt with the composition of the transducers that insert blanks (_) between letters, and accept the language $L(\_L)^+$, where L is the language of all entries in the standard dictionary. Also, candidates within a Levenshtein distance of one are generated. Finally, the Longest Common Subsequence is calculated between the OOV word and each normalization candidate, thus removing candidates whose ratio is below a threshold.

[4]Considering the formation of the digraphs "rr" and "ll" as valid repetitions in the Spanish language.

[5]`https://pypi.python.org/pypi/Unidecode`

[6]This transducer generates other versions of the OOV word by accentuating its vowels (only one vowel is accentuated per version).

[7]The composition is made in the order in which the transducers are stated.

3.3 Candidate Selection

To select the best normalization candidate from the confusion set of an OOV word, contextual information is taken into account. However, because in a context can co-occur several non-standard forms, the selection of the normalization candidates corresponds to the candidates combination that maximizes an objective function. Therefore, the combinations are evaluated against a language model implemented with the Kenlm tool [9], and the one that obtains the highest log probability of sequence of words is selected. The model was estimated from the Spanish Wikipedia corpus.

3.4 Post-processing

Even though the best normalization candidates have been selected, it may still be required a post-processing for the proper capitalization of them. In the Spanish language [17], capital letters are used to differentiate proper nouns from common nouns; however, the case is also required by the punctuation. For proper nouns, their capitalization is selected by the application of the language model. Otherwise, a selected candidate is uppercased if one of the following conditions is satisfied:

1. If the OOV word is in initial position of tweet.
2. If the OOV word is preceded by one of the following punctuation marks: ". ! ?".[8]
3. If the previous token is an ellipsis mark, and the OOV word begins with an uppercase letter.

4. RESOURCES

The system employs a set of lexical resources to suggest normalization candidates for OOV words. The set consists of a dictionary of Spanish standard words, a normalization dictionary, and a gazetteer of proper nouns, which are described below. While the normalization dictionary was entirely handcrafted, the other lexical resources have been built in an automatic way.

4.1 Standard Dictionary

The dictionary of Spanish standard words was built from the FreeLing Spanish dictionary of 556,509 forms. This one was expanded with the entries in the Dirae lexicon[9], and by generating verbal forms of *voseo* [19]. The final dictionary consists of 619,550 standard word forms. Note that the inflected forms of enclitic pronouns, adverbs ending in *-mente*, and diminutives were not added as entries in the dictionary; they are recognized during the process of confusion set generation by applying a set of morphological rules.

4.2 Normalization Dictionary

The normalization dictionary consists of 529 entries that correspond to initialisms, shortenings, and other Internet slang expressions frequently used in Spanish tweets. This resource was mainly built from the most frequent OOV words observed in a development set, for which can be provided their normalization unambiguously. In this way, for each OOV word in the dictionary, its canonical form is included.

[8]The double quotes are used to enclose the punctuation marks.

[9]http://dirae.es/

4.3 Gazetteer of Proper Nouns

The list of proper nouns was built by following the approach in [21]. The Spanish Wikipedia corpus was morphologically analyzed using FreeLing, being the forms categorized as named entities considered as candidates to build the gazetteer. These forms were tokenized and those unigrams whose frequency was greater than 100 and higher than their lowercased variant, and which were not found in the standard dictionary, were taken as secure proper nouns. In this way, a gazetteer of 53,531 unigrams was built.

5. EXPERIMENTS AND EVALUATION

In this section the experimentation with the system to set its parameters and the evaluation of its performance are discussed. To carry out these processes, the resources provided by the organizing committee of the TweetNorm 2013 shared task [2], which are a benchmark for lexical normalization of tweets written in Spanish, have been employed. These resources comprise a set of 937 tweets divided into two collections, i.e., the development corpus (475 tweets) and the test corpus (462 tweets), with 653 and 572 OOV words manually annotated, respectively.[10] The RAE dictionary[11] was taken as reference to determine the standard word forms.

In Section 5.1 the metrics used to evaluate the performance of the system are described. The experimentation conducted on the development corpus to set the parameters of the system, regarding the ability of OOV words detection and the contextual information required to select normalization candidates, is discussed in Section 5.2. Finally, the evaluation of the system on the test set, conceiving it as a whole and by isolating its components, is discussed in Section 5.3.

5.1 Metrics

The detection rate metric evaluates the ability of the system to detect OOV words. The candidate coverage metric [5] measures how many times the confusion set of an OOV word includes the proper correction, regardless of the candidate selection. The standard information retrieval metrics of precision, recall, and F1-score have been also used to evaluate the performance of the system. These five metrics are described below:

$$Detection\ rate = \frac{\sum_{t \epsilon T} \sum_{oov \epsilon OOV'_t} [oov \epsilon OOV_t]}{\sum_{t \epsilon T} |OOV_t|}$$

$$Candidate\ coverage = \frac{\sum_{t \epsilon T} \sum_{oov \epsilon OOV'_t} [corr^t_{oov} \epsilon C^t_{oov}]}{\sum_{t \epsilon T} |OOV'_t|}$$

$$Precision\ (P) = \frac{\sum_{t \epsilon T} \sum_{oov \epsilon OOV'_t} [sel^t_{oov} = corr^t_{oov}]}{\sum_{t \epsilon T} |OOV'_t|}$$

$$Recall\ (R) = \frac{\sum_{t \epsilon T} \sum_{oov \epsilon OOV'_t} [sel^t_{oov} = corr^t_{oov}]}{\sum_{t \epsilon T} |OOV_t|}$$

$$F1\text{-}score\ (F) = \frac{2 \times P \times R}{P + R}$$

[10]At the time of tweets collection retrieval, on July 2015, several tweets had been removed from the Twitter historical data. Therefore, of 1,162 tweets provided, 937 were retrieved by using the Twitter REST APIs.

[11]`http://dle.rae.es/`

Table 1: Performance of the system on the test set with different isolated components. All values are given in percentages

Active components	Candidate coverage	P	R	F1
All	79.65	69.65	69.41	69.53
All – Matching simple rules	68.95	55.96	55.77	55.86
All – Confusion set generation	63.68	61.40	61.19	61.29
All – Phonetic transcription – Suffixes search	80.35	64.39	64.16	64.27
All – Vowels insertion – Edit distance – Split words	74.21	69.30	69.06	69.18
All – Post-processing	72.46	62.11	61.89	62.00

Where,

- T is the collection of tweets, OOV_t the set of OOV words in tweet $t \in T$, and OOV_t' the set of detected OOV words.
- C_{oov}^t is the confusion set of an OOV word, sel_{oov}^t the normalization candidate selected from the confusion set, and $corr_{oov}^t$ the proper correction of the OOV word.

5.2 Setting the System

A critical step of the process of lexical normalization has to do with the ability of the system to detect OOV words. Thus, if several OOV words are not detected, recall could significantly drop. For this reason, two approaches of OOV words detection have been proposed: in the first one, the tokens without analysis by the morphological analyzer of FreeLing are treated as OOV words; in the second one, in addition to the tokens without analysis, the named entities are also treated as OOV words, as it was discussed in Section 3.1. To select between these approaches, several experiments were conducted on the development set. With a detection rate of 98.77%, and 23 percentage points higher than that of the first approach, the second one was selected to detect OOV words.

Likewise, the amount of contextual information required by the candidate selection component was determined. In this way, different orders of the language model were evaluated. As result, the highest precision was obtained by a 3-grams language model, 71.78%, above the 71.32% that both 2- and 4-grams achieve.

5.3 Results and Evaluation

Table 1 shows the performance values of the system on the test set. Here, the system was evaluated by activating all its components; likewise, a further evaluation was conducted by isolating each component in order to determine its contribution to the overall performance. In general, the system achieves a F1-score of 69.53%, with a precision of 69.65% and recall of 69.41%.

Clearly the results show that the greatest room for improvement is in the candidate selection component. The language used in the Spanish Wikipedia corpus, from which the language model was estimated, is characterized by being more formal than that used in Twitter, where predominates a free writing style. Therefore, it should be considered a language model that adapts to informal genres. Despite the prevalence of OOV words in Twitter data, it is not difficult to build a large corpus of tweets with only standard word forms [7]. As future work, it is planned to build a large corpus of tweets from user accounts who, in theory, write correctly, e.g., journalists and mass media.

Table 2: Performance comparison with participating systems in the TweetNorm 2013 shared task

Rank	System	R
1	RAE [15]	78.32%
2	**ours**	**69.41%**
3	Citius-Imaxin [6]	66.43%
4	UPC [1]	65.56%
5	Elhuyar [21]	63.81%

With regard to the contribution of each component to the overall performance, it is observed that the matching simple rules and the post-processing are which contribute the most, thus when these are deactivated, the performance of the system drops significantly. The most complex cases of OOV words are mainly dealt with the phonetic transcription and the suffixes search; instead, deactivating the normalization candidates generation through the vowels insertion, edit distance, and splitting of words, causes a negligible drop in the overall performance.

Finally, the performance of the system was compared with the participating systems in the TweetNorm 2013 shared task. This comparison was made by considering only the 462 tweets of the test set that were retrieved. The best five results sorted by recall are shown in Table 2.[12] For reference, recall average of the 13 participating systems was 56.52%, with the lowest score being 33.93%.

6. CONCLUSIONS

In this paper, a lexical normalization system of tweets written in Spanish was proposed. The system correctly detects OOV words in tweets and suggests normalization candidates that are identical or similar to the graphemes or phonemes that make an OOV word. To select the best normalization candidate for an OOV word, contextual information is taken into account. However, because in a context can co-occur other OOV words, the selection corresponds to the candidates combination that best fits a trigram language model. Although the most of cases the correct normalization of an OOV word is suggested, there is a room for improvement in the candidate selection, which is not properly adapted to the informal genre and the free writing style of Twitter.

As future work, it is planned to build a large corpus of tweets from user accounts who, in theory, write correctly, in order to improve the performance of the candidate selection, and thus adapting it to the genre and language used in Twitter.

[12]Recall was the official metric used to evaluate the performance of the systems in the shared task.

7. ACKNOWLEDGEMENTS

The authors sincerely thank Camilo López and Róbinson Alvarado for their valuable comments and suggestions.

8. REFERENCES

[1] A. Ageno, P. R. Comas, L. Padró, and J. Turmo. The talp-upc approach to tweet-norm 2013. In *Proceedings of the Tweet Normalization Workshop at SEPLN 2013*, September 2013.

[2] I. Alegria, N. Aranberri, P. R. Comas, V. Fresno, P. Gamallo, L. Padró, I. S. Vicente, J. Turmo, and A. Zubiaga. Tweetnorm: a benchmark for lexical normalization of spanish tweets. *Language Resources and Evaluation*, 49(4):883–905, 2015.

[3] K. R. Beesley and L. Karttunen. A gentle introduction. In *Finite State Morphology*. Center for the Study of Language and Information, April 2003.

[4] F. Bravo-Marquez, M. Mendoza, and B. Poblete. Combining strengths, emotions and polarities for boosting twitter sentiment analysis. In *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining*, WISDOM '13, 2013.

[5] J. Cotelo, F. Cruz, J. Troyano, and F. Ortega. A modular approach for lexical normalization applied to Spanish tweets. *Expert Systems with Applications*, 42(10):4743–4754, 2015.

[6] P. Gamallo, M. García, and J. R. Pichel. A method to lexical normalisation of tweets. In *Proceedings of the Tweet Normalization Workshop at SEPLN 2013*, September 2013.

[7] B. Han and T. Baldwin. Lexical normalisation of short text messages: Makn sens a #twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 368–378, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[8] B. Han, P. Cook, and T. Baldwin. Automatically constructing a normalisation dictionary for microblogs. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 421–432, 2012.

[9] K. Heafield. Kenlm: Faster and smaller language model queries. In *Proceedings of the EMNLP 2011 Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland, United Kingdom, July 2011.

[10] M. Hulden. Foma: a finite-state compiler and library. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 29–32. Association for Computational Linguistics, 2009.

[11] R. Li, K. H. Lei, R. Khadiwala, and K. C.-C. Chang. Tedas: A twitter-based event detection and analysis system. In *Data Engineering (ICDE), 2012 IEEE 28th International Conference on*, pages 1273–1276, April 2012.

[12] B. Liu. Sentiment analysis: A multifaceted problem. *IEEE Intelligent Systems*, 25(3):76–80, 2010.

[13] O. Owoputi, B. O'Connor, C. Dyer, K. Gimpel, N. Schneider, and N. A. Smith. Improved part-of-speech tagging for online conversational text with word clusters. In *In Proceedings of NAACL 2013*, 2013.

[14] L. Padró and E. Stanilovsky. Freeling 3.0: Towards wider multilinguality. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, Istanbul, Turkey, May 2012. ELRA.

[15] J. Porta and J. L. Sancho. Word normalization in Twitter using finite-state transducers. In *Proceedings of the Tweet Normalization Workshop at SEPLN 2013*, September 2013.

[16] RAE. Exclusión de *ch* y *ll* del abecedario. `http://www.rae.es/consultas/exclusion-de-ch-y-ll-del-abecedario`. (accessed: October 16, 2015).

[17] RAE. Mayúculas. `http://buscon.rae.es/dpd/srv/search?id=BapzSnotjD6n0vZiTp`. (accessed: October 15, 2015).

[18] RAE. Seseo. `http://lema.rae.es/dpd/srv/search?id=IIUwJDU07D6XC2xEky`. (accessed: November 9, 2015).

[19] RAE. Voseo. `http://lema.rae.es/dpd/srv/search?id=iOTUSehtID6mVONyGX`. (accessed: October 24, 2015).

[20] RAE. Yeísmo. `http://lema.rae.es/dpd/srv/search?id=HK5DEyboyD6iOqnxZu`. (accessed: October 23, 2015).

[21] X. Saralegi and I. S. Vicente. Elhuyar at tweetnorm 2013. In *Proceedings of the Tweet Normalization Workshop at SEPLN 2013*, September 2013.

[22] H. Schoen, D. Gayo-Avello, P. T. Metaxas, E. Mustafaraj, M. Strohmaier, and P. Gloor. The power of prediction with social media. *Internet Research*, 23(5):528–543, 2013.

[23] A. Seshagiri. The languages of twitter users. `http://bits.blogs.nytimes.com/2014/03/09/the-languages-of-twitter-users/`. (accessed: December 4, 2015).

[24] J. Stecyk. Study: Twitter users love mobile apps. `https://blog.twitter.com/2015/study-twitter-users-love-mobile-apps`. (accessed: November 10, 2015).

[25] R. Zacarías. Formación de diminutivos con el sufijo /-ít-/. una propuesta desde la morfología natural. *Anuario de Letras: Lingüística y Filología*, 44:77–103, 2006.